University of Zurich[UZH]

Department of Informatics

University of Zürich
Department of Informatics
Binzmühlestr. 14
CH-8050 Zürich
Phone. +41 44 635 43 11
Fax +41 44 635 68 09
www.ifi.uzh.ch/dbtg

UZH, Dept. of Informatics, Binzmühlestr. 14, CH-8050 Zürich

Prof. Dr. Michael Böhlen
Professor
Phone +41 44 635 43 33
Fax +41 44 635 68 09
boehlen@ifi.uzh.ch

Zürich, 10. November 2023

Bachelor Thesis (18 ECTS)
Database Technology

Topic: Integrating Dimensionality Reduction Computations into MonetDB

There are significant amounts of business data that are maintained in databases. This data must be analyzed to extract actionable business intelligence. The goal of this Bachelor thesis is to extend MonetDB with linear algebra operations that yield results with a schema that depends on the values in relations, implement efficient evaluation techniques, and evaluate the solution in terms of performance and applicability.

Examples of operations that yield results with a schema that depends on the values in relations are matrix transpose (`TRA`), the outer product (`OPD`), and matrix U from a singular value decomposition (`USV`) [2]. Incorporating these operations into MonetDB creates new possibilities for advanced data analyses.

The outer product, for example, can be used for autocovariance computations [6], image compression [7], matrix multiplication accelerator [8] and recommendation computations [4]. Singular value decomposition is used for dimensionality reduction [9], least square computations [3], principal component analysis (PCA) and feature engineering [10].

Example: Consider relation m as follows:

m:

movie	realisticness(f1)	colorfulness(f2)
Oppenheimer	1	-1
Barbie	-1	1
Mission Impossible	0	0

The outer product constructs a vector space with the relations between movie pairs:

OPD(m by movie, m by movie):

a:

movie	Oppenheimer	Barbie	Mission Impossible
Oppenheimer	2	-2	0
Barbie	-2	2	0
Mission Impossible	0	0	0

The SVD decomposition of a yields relation u, d and v:

u:

movie	Oppenheimer	Barbie	Mission Impossible
Oppenheimer	-0.707	0.707	0
Barbie	0.707	0.707	0
Mission Impossible	0	0	1

d:

movie	realisticness(f1)	colorfulness(f2)
Oppenheimer	2	0
Barbie	0	0
Mission Impossible	0	0

v:

c	realisticness(f1)	colorfulness(f2)
realisticness(f1)	-0.707	0.707
colorfulness(f2)	0.707	0.707

According to the analysis of matrix d, there is one dominant singular value. Therefore, we can reduce the dimensionality by linearly transforming the initial data based on the corresponding columns in u and the corresponding rows in v.

Integrating `OPD` and `USV` into MonetDB is difficult because the MonetDB pipeline assumes the schema of the result can be determined from the schema without accessing the data. Since this is not possible for `OPD` and `USV` the processing pipeline of MonetDB must be modified to allow query plans with partial schema information. At the physical level new methods must be designed that make it possible to efficiently deal with partially instantiated query plans in column store systems [1, 5].

The tasks of the BSc thesis are described below. At the end a carefully worked out BSc thesis that describes the solutions to these tasks must be handed in. The results shall be presented at a DBTG meeting.

Task 1: Implementation of OPD and SVD

Implement and integrate the outer product and SVD operations into MonetDB. Particular attention must be paid to the handling of schema information. Since the schema of an outer product result relation cannot be derived from existing schema information and the query a new evaluation approach must be implemented. The integration of the outer product operation into the query evaluation pipeline of MonetDB must be designed so that the outer product can be used in nested sequences of operations. To achieve this the following elements of MonetDB must be studied and extended:

- overview, deployment - 1 week
- SQL syntax, symbol tree (sql_parser.y) - 1 week

- relation tree (rel_select.c) - 1 week
- statement tree (rel_bin.c) - 2 week
- execution backend (batcalc.c, gdk_calc.c) - 3 weeks
- evaluation - 2 weeks

Task 2: (2 weeks) Study, describe and implement application examples from statistics, machine learning, and data science that use the outer product. Specifically, SVD shall be used to perform an effective dimensionality reduction of the data.

References

[1] Peter A Boncz, Marcin Zukowski, and Niels Nes. Monetdb/x100: Hyper-pipelining query execution. In *Cidr*, volume 5, pages 225–237, 2005.

[2] Oksana Dolmatova, Nikolaus Augsten, and Michael H Böhlen. A relational matrix algebra and its implementation in a column store. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pages 2573–2587, 2020.

[3] Gene H Golub and Christian Reinsch. Singular value decomposition and least squares solutions. In *Handbook for Automatic Computation: Volume II: Linear Algebra*, pages 134–151. Springer, 1971.

[4] Xiangnan He, Xiaoyu Du, Xiang Wang, Feng Tian, Jinhui Tang, and Tat-Seng Chua. Outer product-based neural collaborative filtering. *arXiv preprint arXiv:1808.03912*, 2018.

[5] Milena G Ivanova, Martin L Kersten, Niels J Nes, and Romulo AP Gonçalves. An architecture for recycling intermediates in a column-store. *ACM Transactions on Database Systems (TODS)*, 35(4):1–43, 2010.

[6] Brian J Odelson, Murali R Rajamani, and James B Rawlings. A new autocovariance least-squares method for estimating noise covariances. *Automatica*, 42(2):303–308, 2006.

[7] Dianne O'Leary and Shmuel Peleg. Digital image compression by outer product expansion. *IEEE Transactions on Communications*, 31(3):441–444, 1983.

[8] Subhankar Pal, Jonathan Beaumont, Dong-Hyeon Park, Aporva Amarnath, Siying Feng, Chaitali Chakrabarti, Hun-Seok Kim, David Blaauw, Trevor Mudge, and Ronald Dreslinski. Outerspace: An outer product based sparse matrix multiplication accelerator. In *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 724–736. IEEE, 2018.

[9] KV Ravi Kanth, Divyakant Agrawal, and Ambuj Singh. Dimensionality reduction for similarity searching in dynamic databases. *ACM SIGMOD Record*, 27(2):166–176, 1998.

[10] Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha. Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, pages 91–109. Springer, 2003.

Supervisor: Xinyu Zhu (xinyu.zhu@uzh.ch)

Start date: November 15, 2023

End date: May 15, 2024

University of Zurich
Department of Informatics

Prof. Dr. Michael Böhlen
Professor